

Calculating the Performance Gain Due to Improved Predictive Validity

D. R. Dlvgl

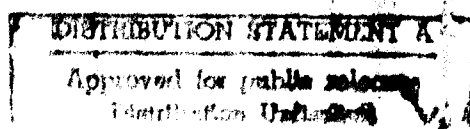
AD-A230 732



A Division of Hudson Institute

CENTER FOR NAVAL ANALYSES

1101 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268



1 1 0 072

APPROVED FOR RELEASE BY NSA ON 08-11-2013

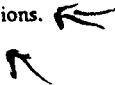
Work conducted under contract N50014-87-C-0001

This Research Memorandum represents the best opinion of CNA at the time of issue. It does not necessarily represent the opinion of the Department of the Navy

REPORT DOCUMENTATION PAGE

Form Approved
OPM No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources gathering and maintaining the data needed, and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Information and Regulatory Affairs, Office of Management and Budget, Washington, DC 20503.

1. AGENCY USE ONLY (Leave Blank)		2. REPORT DATE April 1990	3. REPORT TYPE AND DATES COVERED Final
4. TITLE AND SUBTITLE Calculating the Performance Gain Due to Improved Predictive Validity		5. FUNDING NUMBERS C - N00014-87-C-0001 PE - 65153M PR - C0031	
6. AUTHOR(S) D.R. Divgi			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Center for Naval Analyses 4401 Ford Avenue Alexandria, Virginia 22302-0268		8. PERFORMING ORGANIZATION REPORT NUMBER CRM 89-254	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Commanding General Marine Corps Combat Development Command (WF 13F) Studies and Analyses Branch Quantico, Virginia 22134		10. SPONSORING/MONITORING AGENCY REPORT NUMBER	
11. SUPPLEMENTARY NOTES			
12a. DISTRIBUTION/AVAILABILITY STATEMENT Approved for Public Release; Distribution Unlimited		12b. DISTRIBUTION CODE	
13. ABSTRACT (Maximum 200 words) Composites of test scores are used in selection and classification of enlisted personnel. If a selection composite is replaced by a new composite with higher predictive validity, mean performance of the recruits increases. Formulas for calculating the performance gain require simplifying assumptions. This research memorandum shows with an example that the formulas are very sensitive to violations of these assumptions and hence are not dependable enough for use in operational decisions. 			
14. SUBJECT TERMS ASVAB (armed services vocational aptitude battery), Calculations, CAT (computerized adaptive testing), Performance (human), Personnel selection, Mathematical prediction, Regression analysis, Scoring, Test methods, Validation		15. NUMBER OF PAGES 12	
		16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT CPR	18. SECURITY CLASSIFICATION OF THIS PAGE CPR	19. SECURITY CLASSIFICATION OF ABSTRACT CPR	20. LIMITATION OF ABSTRACT SAR

NSN 7540-01-280-5500

Standard Form 298, (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
299-01



CENTER FOR NAVAL ANALYSES

A Division of Hudson Institute 4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268 • (703) 824-2000

8 May 1990

MEMORANDUM FOR DISTRIBUTION LIST

Subj: Center for Naval Analyses Research Memorandum 89-254

Encl: (1) CNA Research Memorandum 89-254, *Calculating the Performance Gain Due to Improved Predictive Validity*, by D.R. Divgi, Unclassified, April 1990

1. Enclosure (1) is forwarded as a matter of possible interest.
2. Composites of test scores are used in selection and classification of enlisted personnel. If a selection composite is replaced by a new composite with higher predictive validity, mean performance of the recruits increases. Formulas for calculating the performance gain require simplifying assumptions. This research memorandum shows with an example that the formulas are very sensitive to violations of these assumptions and hence are not dependable enough for use in operational decisions.

Lewis R. Cabe
Director
Manpower and Training Program

Distribution List:
Reverse page

Accession For	
NTIS GRA&I	<input checked="" type="checkbox"/>
DTIC TAB	<input checked="" type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	



Subj: Center for Naval Analyses Research Memorandum 89-254

Distribution List

SNDL

A1	ASSTSECNAV MRA
A1	DASN - MANPOWER
A2A	CNR
A6	HQMC MPR
	Attn: M
	Attn: MP
	Attn: MR
	Attn: MA (2 copies)
	Attn: MPP-54
FF38	USNA
	Attn: Nimitz Library
FF42	NAVPGSCOL
FF44	NAVWARCOL
	Attn: E-111
FJA1	COMNAVMIIPERSCOM
FJB1	COMNAVCRUITCOM
FKQ6D	NAVPERSRANDCEN
	Attn: Technical Director (Code 01)
	Attn: Director, Testing Systems (Code 13)
	Attn: Technical Library
	Attn: Director, Personnel Systems (Code 12)
	Attn: CAT/ASVAB PMO
	Attn: Manpower Systems (Code 11)
FT1	CNET
V12	CGMCCDC
	Attn: Training and Education Center
	Attn: Warfighting Center (WF-13F)

OPNAV

OP-01SA

OP-11B

OP-136

OTHER

Joint Service Selection and Classification Working Group (13 copies)

Defense Advisory Committee on Military Personnel Testing (8 copies)

Calculating the Performance Gain Due to Improved Predictive Validity

D. R. Divgi

Force Structure and Acquisition Division



A Division of Hudson Institute

CENTER FOR NAVAL ANALYSES

4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268

ABSTRACT

Composites of test scores are used in selection and classification of enlisted personnel. If a selection composite is replaced by a new composite with higher predictive validity, mean performance of the recruits increases. Formulas for calculating the performance gain require simplifying assumptions. This research memorandum shows with an example that the formulas are very sensitive to violations of these assumptions and hence are not dependable enough for use in operational decisions.

EXECUTIVE SUMMARY

The Armed Services Vocational Aptitude Battery (ASVAB) is used for selection and classification of enlisted personnel. The ASVAB is useful to DOD because of its predictive validity (i.e., its ability to predict performance on the job). A computerized adaptive testing (CAT) version of the ASVAB has been developed. Through computerized testing, value of the ASVAB may be increased by adding new tests that cannot be administered at present. The utility of adding such tests has been estimated in a cost/benefit analysis of the CAT-ASVAB project [2].

The formulas used in calculating the benefit of new predictors require simplifying assumptions. Such assumptions are bound to be violated to some extent in reality. If a formula is sensitive to violations of its assumptions, the actual benefit may be quite different from the value given by the formula.

This research memorandum uses data from the Marine Corps Job Performance Measurement Project and from military applicants tested in late 1984. It compares benefits calculated in five different ways. The results show that the calculated benefit may halve or double from one formula to another. Thus, the benefit estimate depends strongly on how it is calculated. Such unstable estimates are not useful in making operational decisions.

CONTENTS

	Page
Introduction	1
Calculating Performance Gain	1
Data for Illustration	2
Incremental Validity	3
Multiple Regression	4
Interpretation	5
Conclusions	7
References	7
Appendix	A-1

INTRODUCTION

The Armed Services Vocational Aptitude Battery (ASVAB) is used for selection and classification of enlisted personnel. It contains ten subtests--General Science (GS), Arithmetic Reasoning (AR), Word Knowledge (WK), Paragraph Comprehension (PC), Numerical Operations (NO), Coding Speed (CS), Auto and Shop Information (AS), Mathematics Knowledge (MK), Mechanical Comprehension (MC), and Electronics Information (EI). An eleventh subtest--Verbal (VE)--is defined as the sum of WK and PC. Standard scores rather than raw scores on the subtests are used in all decisions based on the ASVAB. Standard scores are integers from 20 to 80, with mean 50 and standard deviation 10 in the 1980 reference population [1]. Standard scores on subtests are combined into the Armed Forces Qualification Test (AFQT) score, which is the same for all services, and into occupational composites that vary from one service to another. The AFQT is the primary score for selection of an applicant for enlistment, while composite scores are used to classify a recruit into one of the available military occupational specialties (MOSs).

A computerized adaptive testing (CAT) version of the ASVAB has been developed. In CAT, a computer program selects items for an examinee on the basis of available information about the examinee's ability. Thus, a capable examinee's time is not wasted on easy items nor a below-average examinee's time on difficult items. As a result, CAT can achieve as much precision as the conventional paper-pencil (PP) version of a test with fewer items. On the average, the CAT-ASVAB takes about half as long as the PP-ASVAB.

The ASVAB is useful to DOD because of its predictive validity (i.e., its ability to predict performance on the job). Recruits selected using the ASVAB perform better than those selected at random. The value of the ASVAB will increase if it is improved by adding new tests that measure traits such as perceptual and psychomotor abilities. The utility of adding such tests has been estimated in a cost/benefit of the CAT-ASVAB project ([2], tab E) using the "Cronbach-Gleser formula." One can derive a number of such formulas that differ in the number of simplifying assumptions required. These assumptions, although reasonable, are likely to be wrong to some extent. This research memorandum demonstrates the sensitivity of these formulas to violations of their assumptions. When the additional validity due to the new tests is small, the effect of departure from the assumptions may be of the same size as the utility being calculated.

CALCULATING PERFORMANCE GAIN

A utility analysis attempts to estimate the performance gain that will result from replacing one composite with a different, more valid composite in a future population of applicants. This analysis is based on information from one or more validity studies relating test scores to some measure of performance. In such a study, correlations between ASVAB subtests and the performance measure are calculated from data on a sample of enlisted personnel. Enlisted personnel have been selected previously using their ASVAB scores. As a result, their scores have a smaller spread than scores for the national population. Therefore, the

sample statistics are adjusted for range restriction. The corrected means, standard deviations, and correlations describe the results that would have been obtained if performance could be measured in the entire national population. These values can be used to calculate the corrected validity (i.e., the correlation with performance in the national population) of any composite score of interest.

The central problem in utility analysis is to apply this knowledge to an unknown applicant population of the future. Because score distributions in such a population are unknown, one must rely on assumptions. Different assumptions lead to different formulas. If all assumptions and hence all formulas are correct, they will yield almost the same value for the performance gain. This paper shows that this does not happen in a large sample of applicants tested in 1984, which means that the simplifying assumptions are incorrect in this sample. The paper examines how the calculated performance gain changes as more and more information from the sample is used.

DATA FOR ILLUSTRATION

A realistic example for calculating performance gains is provided by the recent change in the composition of the AFQT. Until 31 December 1988, the AFQT contained subtests AR, WK, PC, and NO. This will be referred to as the old AFQT. Its raw score is given by

$$\text{OLD_RAW} = \text{AR} + \text{WK} + \text{PC} + \text{NO}/2 \quad (1)$$

The new AFQT, implemented on 1 January 1989, contains MK instead of NO, and uses standard scores instead of raw scores. Thus, the sum of standard scores is

$$\text{NEW_SSS} = 2 \text{ SVE} + \text{SAR} + \text{SMK} \quad (2)$$

where SVE is the standard score on VE, and so on. These scores were standardized in this study so as to have a mean of 0 and a standard deviation of 100 in the reference population [1]. The standardized scores will be referred to as OLD and NEW.

Data from the Marine Corps Job Performance Measurement (JPM) provided scores on a hands-on performance test (HOPT). HOPT scores were standardized to have a standard deviation of 10 in the reference population. The sample from MOS 0351 (Assaultman) was used because it provided the highest incremental validity for the new AFQT over the old AFQT. After eliminating the effect of time-in-service, the predicted HOPT score was given by

$$\begin{aligned} \text{HOPT_PRED} = & 7.9104 + 0.0047 \text{ SGS} + 0.1293 \text{ SAR} + 0.0053 \text{ SWK} - 0.0429 \text{ SPC} \\ & - 0.1564 \text{ SNO} + 0.0901 \text{ SCS} + 0.1066 \text{ SAS} + 0.1667 \text{ SMK} \\ & + 0.1291 \text{ SMC} + 0.2209 \text{ SEI} \quad (3) \end{aligned}$$

The standard error of estimate was 8.088.

This regression equation was assumed to be valid in all populations. Given this assumption, one can calculate the correlation of HOPT

in any population with any ASVAB subtest or composite, provided the distribution of ASVAB subtest scores in that population is known. The validities of OLD and NEW in the reference population turned out to be 0.4475 and 0.4841, respectively. Thus, the increase in predictive validity was 0.0366.

The applicant sample consisted of those who took Form 15c in the Initial Operational Test and Evaluation (IOT&E) of ASVAB forms 11/12/13 in 1984. It has been shown by Maier and Hiatt [3] that, by this time, scores on the speeded subtests suffered from score inflation relative to the 1980 reference population. Therefore, NO and CS scores were adjusted for inflation using the equating approach of Maier and Hiatt. The sample size was 15,065. The cut scores were set so that about 90 percent of the sample would be selected for military service. The minimum acceptable scores turned out to be -107 on OLD and -106 on NEW. The corresponding numbers of selected applicants were 13,578 and 13,561. All calculations are made for illustration only. Random errors of sample statistics are of no interest. Therefore, the distinction between samples and populations will be ignored.

INCREMENTAL VALIDITY

The phrase "incremental validity" is used frequently in connection with new tests. It often means the increase in multiple correlation when the new test is added to the ASVAB. This meaning, however, is irrelevant to a utility analysis because DOD uses composite scores, not multiple regression, in selection and classification. The discussion in this memorandum will use only the composites OLD and NEW. A proper analysis must also take into account the distinction between selection and classification, which makes the analysis very difficult. The calculations in the CAT-ASVAB cost/benefit analysis [2] considered selection only and so will the formulas in this paper. "Incremental validity" will mean the increase in the correlation with HOPT on replacing OLD with NEW.

Any correlation depends on the spread of scores in the population, and hence cannot be assumed to be the same in the reference and applicant populations. It seems reasonable, however, to assume that the regression of performance on the composite remains the same. Given this assumption, and the variances of the composite in the two populations, one can calculate its validity in the applicant population. Because nothing is known about the applicant population of the future, a 1984 applicant sample is used as a substitute. Even in this sample, if a new test were really being evaluated, nothing would be known about the new composite. So, at first, only information about OLD is used.

The standard deviation of OLD in the applicant sample is 78.32. Assuming that NEW has the same spread, validities of OLD and NEW are 0.3649 and 0.3976. As was to be expected, due to the decrease in standard deviation from 100 to 78.32, these validities are lower than in the reference population. The incremental validity in the applicant group is 0.0327. The same theory yields 9.605 as the standard deviation of HOPT among applicants.

Assume, as is done implicitly in [2], that composite distributions among applicants are normal. The normalized z-score corresponding to the selection ratio of 90 percent is -1.28, and the height of the ordinate at this score is 0.176. Therefore, using equation 1.10 from Cronbach and Gleser [4, p. 308], the performance gain per recruit is

$$G1 = 0.0327 (0.176/0.9) (9.605) = 0.0614 \quad . \quad (4)$$

Now, discard the assumption of normality, and use the actual means of OLD among all applicants and among those selected. These are -0.915 and 15.508, respectively. It is assumed that distributions of NEW and OLD have the same shape. When the factor (0.176/0.9) is replaced by the difference between means divided by the standard deviation of OLD, the gain per recruit is

$$G2 = 0.0327 (15.508 + 0.915)/78.32 (9.605) = 0.0659 \quad . \quad (5)$$

Thus, the result does not change much when the assumption of normality is dropped.

Now, drop the assumption that distributions of OLD and NEW are similar, and use the actual sample statistics of NEW. The standard deviation is 78.91, and the adjusted validity of NEW is 0.4001. Hence, the incremental validity is 0.0352. The means among total and selected groups are 0.981 and 16.835. Cronbach and Gleser's equation 1.10 now yields

$$\begin{aligned} G3 &= [0.4001(16.835 - 0.981)/78.91 - 0.3649(15.508 + 0.915)/78.32](9.605) \\ &= [0.080385 - 0.076516] (9.605) = 0.0372 \quad . \quad (6) \end{aligned}$$

Thus, the difference between the distributions of NEW and OLD is enough to cut the performance gain almost by half. This happens even though the two composites share three of their four subtests and the incremental validity is higher than in G2.

MULTIPLE REGRESSION

When the stronger univariate assumptions are discarded and only the multiple regression equation 3 is used instead, the standard deviation of HOPT is 9.682, and the validities of OLD and NEW are 0.3797 and 0.4390. With these validities in the formula above, the gain is

$$\begin{aligned} G4 &= [0.4390 (15.854/78.91) - 0.3797 (16.423/78.32)] (9.682) \\ &= [0.088201 - 0.079620] (9.682) = 0.0831 \quad . \quad (7) \end{aligned}$$

This is substantially higher than any of the previous estimates. The increase is due to the fact that the true incremental validity among applicants is 0.0593 rather than 0.0327 or 0.0352.

Now, ignore the composites altogether, and use the full multiple regression to calculate the gain. Mean HOPT is 41.6058 among those

selected using OLD and 41.7215 among those selected using NEW. Thus, the actual performance gain per recruit is

$$G5 - 41.7215 - 41.6058 = 0.1157 \quad . \quad (8)$$

This is higher even than G4, about twice as large as G1 and G2, and over three times G3.

INTERPRETATION

Old and new AFQT have three subtests in common: AR, WK, and PC. It is reasonable to assume that distributions of their scores have the same shape. Yet this assumption is wrong, and its effect is that gain G2 is about twice as big as G3. This issue, however, is not very important because assumptions about the shapes of distributions are not central to discussions of validity and utility in selection.

The distinction between estimates G1 to G3 on the one hand and G4 and G5 on the other deserves careful attention. The former are based on simple regression on a single composite at a time, while the latter use multiple regression on all subtests.

In discussions of validity studies, it is customary to correct correlations to the reference population, and then evaluate composite validities obtained from the corrected correlation matrix. If the correlation (or covariance) matrix is the basis of all calculations, it is impossible to detect any nonlinearity in the regression of the criterion on the composite. It is important to note that linear regression on subtests (or even on the two composites OLD and NEW) does not guarantee linear regression on a single composite. In the reference population, when the square of OLD was included in the regression equation, it explained more than 5 percent as much variance as was explained by the linear term. Usually, a contribution of this size is safe to ignore. When one is studying the value of incremental validity, however, the quantities of interest are themselves quite small. Therefore, one must not ignore other small influences such as quadratic terms in regression equations.

Emphasis on the reference population tends to divert attention away from the various ways in which it differs from the applicant population. It is reasonable to assume that the regression of HOPT on a composite is the same in both populations, even if correlations differ because variances are different. This assumption leads to the univariate adjustments of validity used in estimates G1 to G3. As has been seen, the assumption is incorrect. In the present example, the actual validities in the applicant group turned out to be higher than the simple estimates, but they might be lower in another situation. Therefore, given the incremental validity of a new composite in the reference population, one does not know what it will be among applicants. For utility analysis, validity in the applicant population is what matters.

Even within the applicant sample, performance gain increased by over a third from G4 to G5. To understand why a small quadratic term

can have such an effect, it is instructive to see where the performance gain comes from. Even though the formulas express the gain in terms of mean scores, the actual gain is not distributed throughout the entire selected group. The gain comes from the superior performance of those who qualify on NEW but not on OLD, over those who qualify on OLD but not on NEW. When the correlation between OLD and NEW is high, these two groups contain only a small fraction of the total applicant sample. The correlation between old and new AFQTs in the 1984 applicant sample is 0.953. Therefore, most applicants who meet the OLD requirement also meet the NEW requirement.

Of the total of 15,065 applicants, 305 qualify on OLD but not on NEW, with mean OLD, NEW and HOPT scores of -88.6, -119.1, and 42.83. The corresponding means are -121.6, -88.4, and 47.79 for the 288 applicants who qualify on NEW but not on OLD. The entire performance gain due to the change in the AFQT comes from the mean difference $47.79 - 42.83 = 4.96$ between these two groups, each of which contains only about 2 percent of the total sample. Both subgroups are near the low end of the score distribution. The means of OLD and NEW in these subgroups differ by over 30 points; in the total sample, the means are almost equal and the standard deviation of the difference is 24. Thus, the applicants in these subgroups have highly unusual patterns of scores. One cannot be confident that relationships that hold in the total sample are valid in these subgroups as well. Influences such as quadratic terms in regression, whose effects appear small in the total sample, can have a major effect on the mean HOPT in these subgroups.

CONCLUSIONS

Calculations in the 1984 IOT&E sample show that the performance gain formulas are very sensitive to small violations of their assumptions. The gain dropped by a half from G2 to G3 because the distributions of OLD and NEW differ in shape, even though only one of four subtests has been changed. It more than doubles from G3 to G4 because simple regressions on the composites are different in the reference and applicant populations. It increases from G4 to G5 by over a third because of small nonlinear components in the regressions of HOPT on OLD and NEW. If the changes from G3 to G5 had been negative instead of positive, use of NEW would have lowered mean performance despite the increased validity in the reference population. The appendix illustrates this possibility with simulated data.

The results in this study came from applying one regression equation (equation 3) to one large applicant sample. This does not weaken the conclusions. A single example suffices to show what can happen, and thus undermines confidence in simple formulas. All that is required is for the regression equation and the data set to be realistic. (Confidence in the formulas becomes even weaker if one rejects the assumption that equation 3 holds for all applicants, including those who qualify on one composite but not the other.) Therefore, the benefit estimates in tab E of the CAT-ASVAB cost/benefit analysis [2], based on the Cronbach-Gleser formula, are not dependable enough to be useful in making operational decisions.

REFERENCES

- [1] CNA Report 116, *The ASVAB Score Scales: 1980 and World War II*, by Milton H. Maier & William H. Sims, Jul 1986 (94011600)¹
- [2] Automated Science Group & CACI, Inc.-Federal, *CAT-ASVAB Program: Concept of Operation and Cost/Benefit Analysis*, Mar 1988
- [3] CNA Research Memorandum 86-228, *Evaluating the Appropriateness of the Numerical Operations and Math Knowledge Subtests in the AFQT*, by Milton H. Maier and Catherine M. Hiatt, Nov 1986 (27860228)
- [4] L. J. Cronbach and G. C. Gleser. *Psychological Tests and Personnel Decisions*. Urbana, Illinois: University of Illinois, 1965

1. The numbers in parentheses are CNA internal control numbers.

APPENDIX

REDUCED PERFORMANCE DUE TO INCREASED VALIDITY

APPENDIX
REDUCED PERFORMANCE DUE TO INCREASED VALIDITY

The following simulation was performed to show that mean performance can indeed go down when the predictive validity of the composite is increased. Simulated OLD scores were standard normal variates multiplied by 100. Correlated normal scores were generated with the equation

$$X = 0.95 \text{ OLD} + E ,$$

where E was normal with a mean of 0 and a standard deviation of 30. These were converted into NEW scores using

$$\text{NEW} = X (1 + X/500) - 20 .$$

These NEW scores had a positively skewed distribution with skewness of 1.1. For each examinee, an HOPT score was computed as

$$\text{HOPT} = 50 + 0.025 \text{ OLD} + 0.035 \text{ NEW} + E' ,$$

where E' was normal with a mean of 0 and a standard deviation of 8 so that the standard deviation of HOPT was 10. The number of simulated examinees was 20,000.

The validities of OLD and NEW were found to be 0.584 and 0.593, so that NEW had higher predictive validity. HOPT was regressed on OLD and on NEW, with squares of the composite scores included as predictors. For OLD, the quadratic term explained 2.35 percent as much variance as the linear term did. For NEW, this percentage was only 0.75. Note that these nonlinear effects were found even though true multiple regression on the two composites was strictly linear.

As in the 1984 sample, each examinee was selected or rejected using OLD and using NEW with a selection ratio of 90 percent. Mean HOPT scores were computed for those selected with OLD but not with NEW, and vice versa. Mean HOPT values were 43.08 for those selected with OLD and 42.76 for those selected with NEW. Although the difference between these numbers is small, the important point is that using OLD yields higher performance even though NEW has higher validity.